# Assignment 03

Michael Cuesta

2/4/2020

```
library(tidyverse)
library(scales)
install.packages("ggplot2")
```

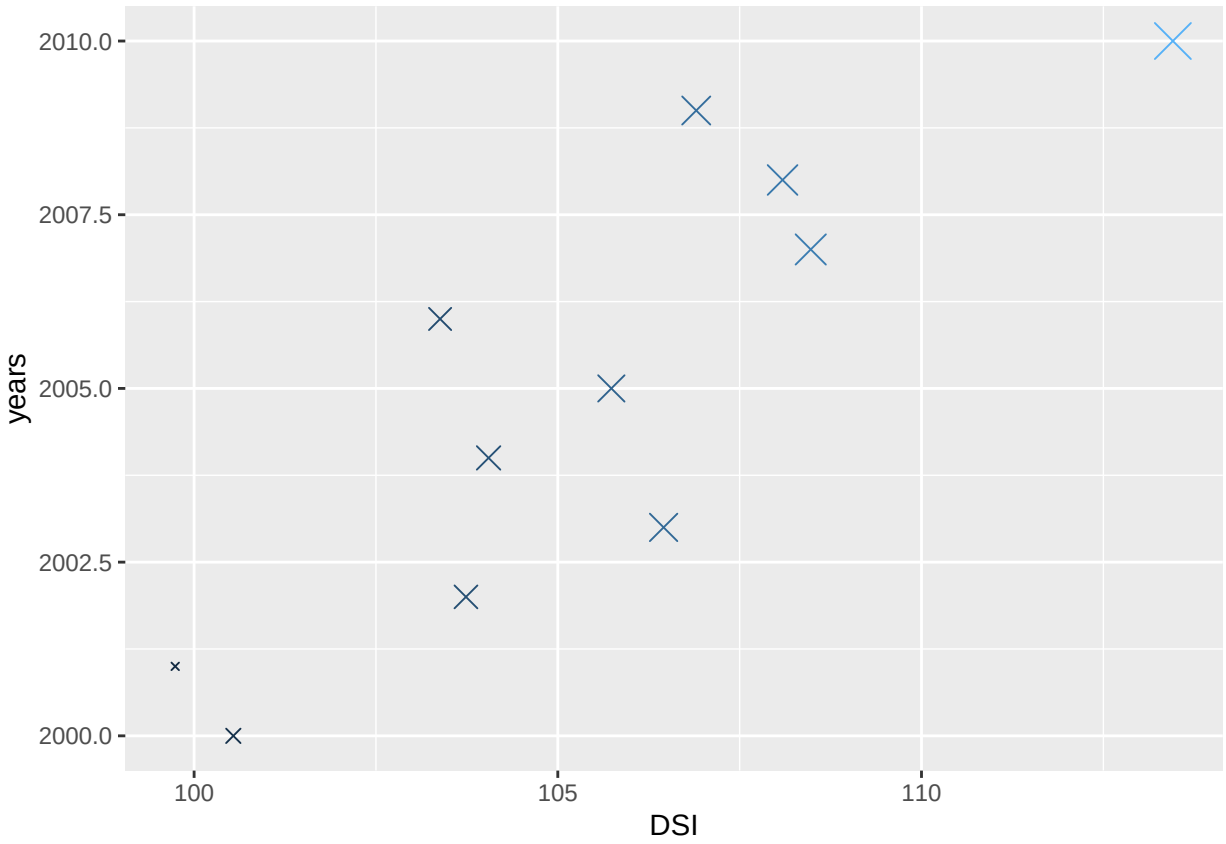
## Question 1

**Using dataset 1 to replicate plot1.pdf. Here you will see that "years" is on the y-axis and "DSI" is on the x-axis.**

```
datset1 <- read_csv("data/datset1.csv")
```

```
## Parsed with column specification:
## cols(
##   years = col_double(),
##   DSI = col_double()
## )
```

```
ggplot(datset1, aes(DSI,years)) +
  geom_point(aes(size=DSI,color=DSI),shape=4, show.legend = FALSE)
```




The graph above is a replication of plot1.

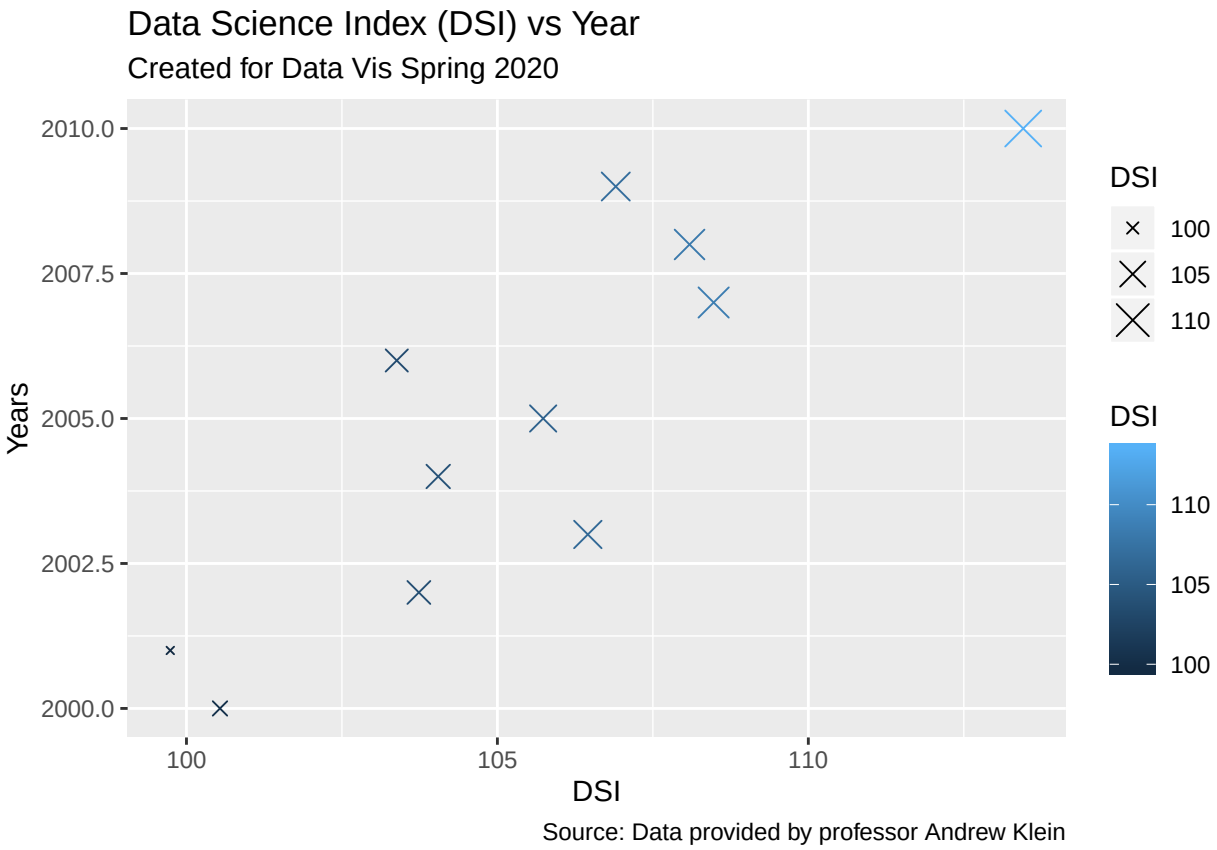
# Question 2

### Improve the plot as you see fit (if any). Note any changes you made.

Below are the improvements made from my replication of plot1 in question one. I kept the points as X's that change in size and color from left to right. I also added a formal title, subtitle, caption, and made the Y-axis "year" into "Year". Legend was added to account for color and size of point.

```
Data1a <- ggplot(datset1, aes(DSI,years)) +
  geom_point(aes(size=DSI,color=DSI),shape=4) +
  labs(x = "DSI", y = "Years", title = "Data Science Index (DSI) vs Year", subtitle = "Created for Data

Data1a
```




## Question 3

### Use the same plotting code from Question 2 and apply it to Dataset 2 - Describe any issues that arise from using the expanded dataset 2

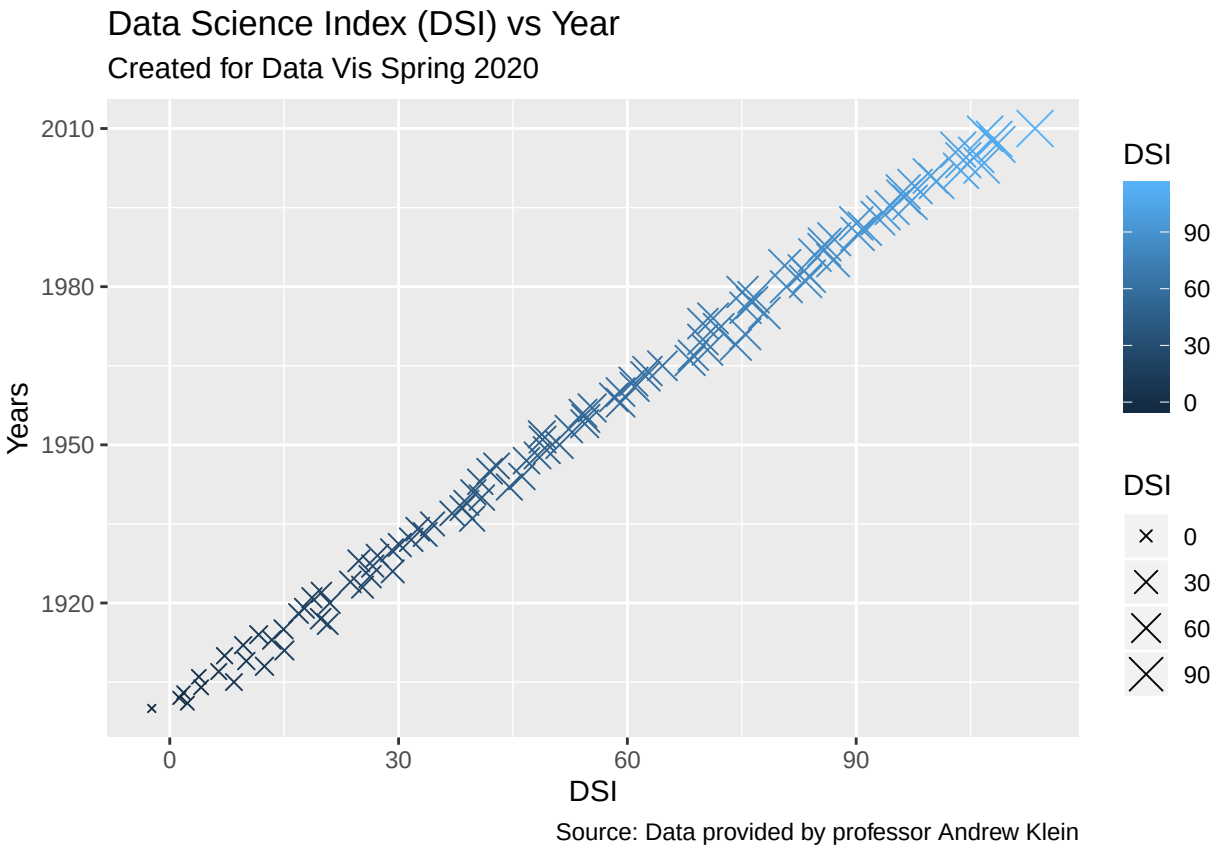
```
datset2 <- read_csv("data/datset2.csv")
```

```
## Parsed with column specification:
## cols(
##   years = col_double(),
##   DSI = col_double()
## )
```

Below is the result of using the exact same code from Question 2 on datset2. When mapped this way datset2 appears to be a very positive linear model. As the year's increase, the Data Science Index also increases. There seems to be no apparent issues. I would, however, revert to just points rather than X's for this graph.

```
Data2a <- ggplot(datset2, aes(DSI,years)) +
  geom_point(aes(size=DSI,color=DSI), shape=4) +
  labs(x = "DSI", y = "Years", title = "Data Science Index (DSI) vs Year", subtitle = "Created for Data


Data2a
```

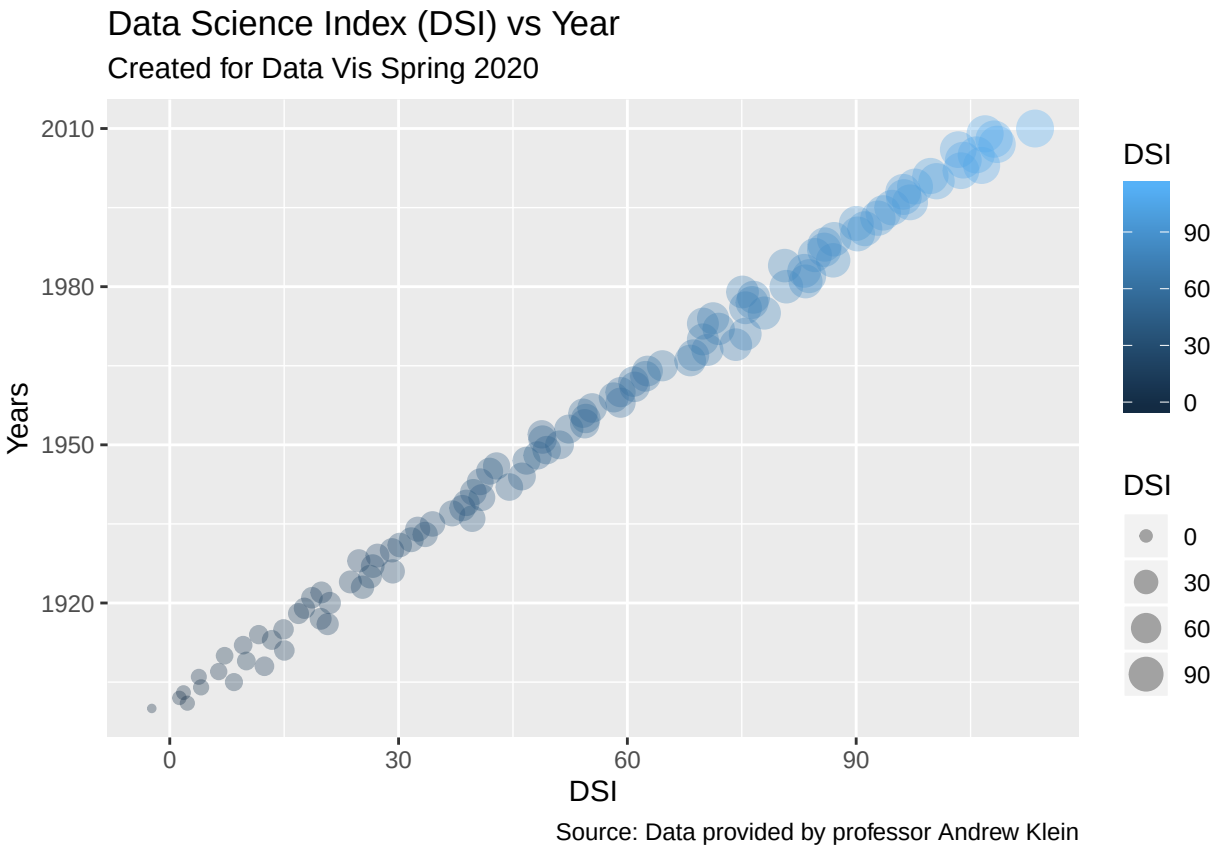



## Question 4

**Make any necessary improvements to your plotting code and create an improved graph (if necessary).**

```
Data2a <- ggplot(datset2, aes(DSI,years)) +
  geom_point(aes(size=DSI,color=DSI), alpha=1/3) +
  labs(x = "DSI", y = "Years", title = "Data Science Index (DSI) vs Year", subtitle = "Created for Data


Data2a
```




For improvements, I removed the X's and added points back. The points now have transparency of 1/3 to allow the viewer to see through any overplotting. Everything else remains the same.

## Question 5

**Dataset 3 - Use ggplot2 to create a graphic using all the variables. Note you may want to consider re-shaping the data before plotting. The values represent GDP per Capita for each country.**

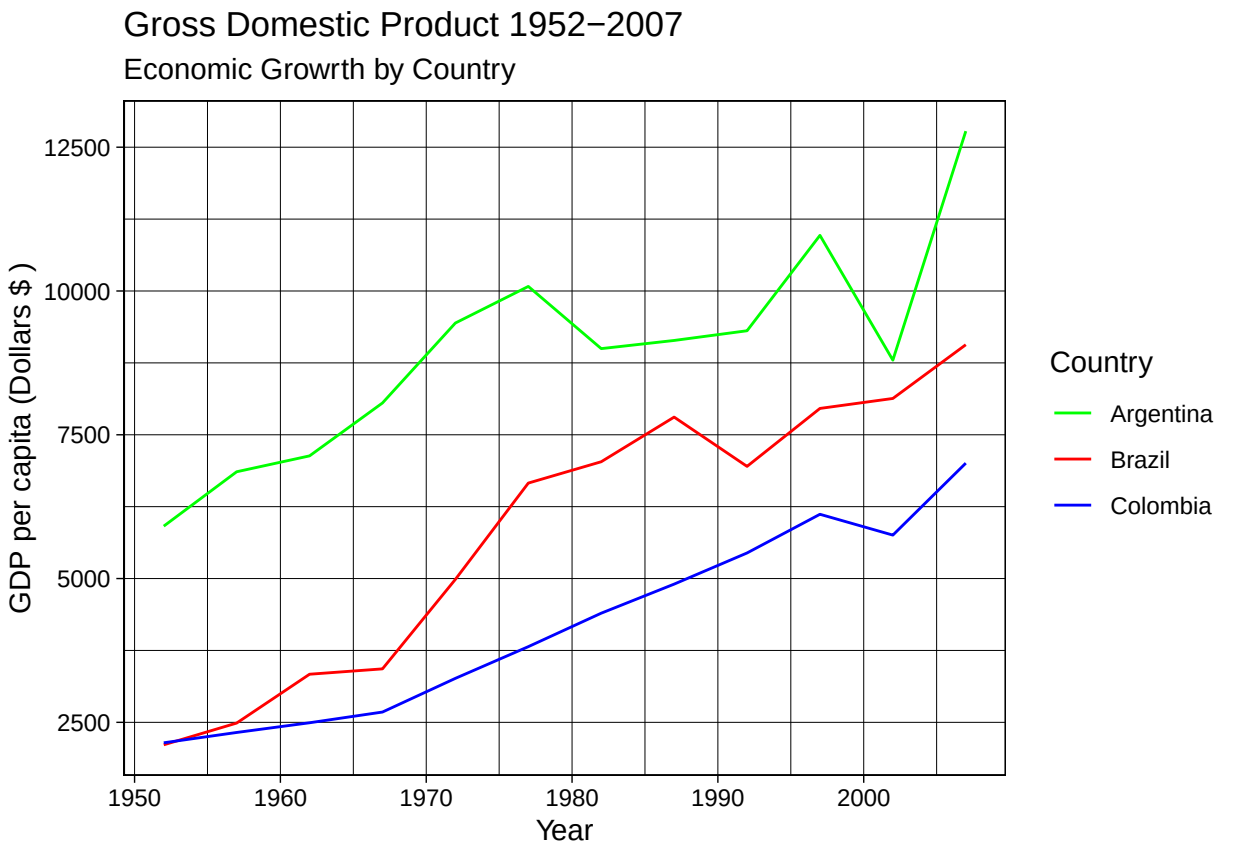
```
datset3 <- read_csv("data/datset3.csv")
```

```
## Parsed with column specification:
## cols(
##   year = col_double(),
##   Argentina = col_double(),
##   Brazil = col_double(),
##   Colombia = col_double()
## )
```

```
Dataset3 <- gather(datset3, "Country", "GDP", 2:4)
```

```
ggplot(Dataset3, aes(year,GDP, color=Country)) +
  theme(plot.background = element_blank())+
  scale_y_continuous(name = "GDP per capita (Dollars $ )") +
  scale_x_continuous(name = "Year") +
```

```
  labs(title= "Gross Domestic Product 1952-2007",
       subtitle="Economic Growrth by Country")+
    geom_line()+
    scale_color_manual(values = c("green", "red", "blue"), guide= guide_legend(title="Country"))+
    theme_linedraw()+
    theme(panel.grid.major = element_line(color = "black"))
```




## Question 6

**Dataset 4 - Use ggplot2 to create a graphic to explain the relationship between the two variables. You may use one of the methods described in class, in the ggplot2 book, or another method of your choosing. Explain why you chose the method you did and the pros and cons are of the method that you chose.**

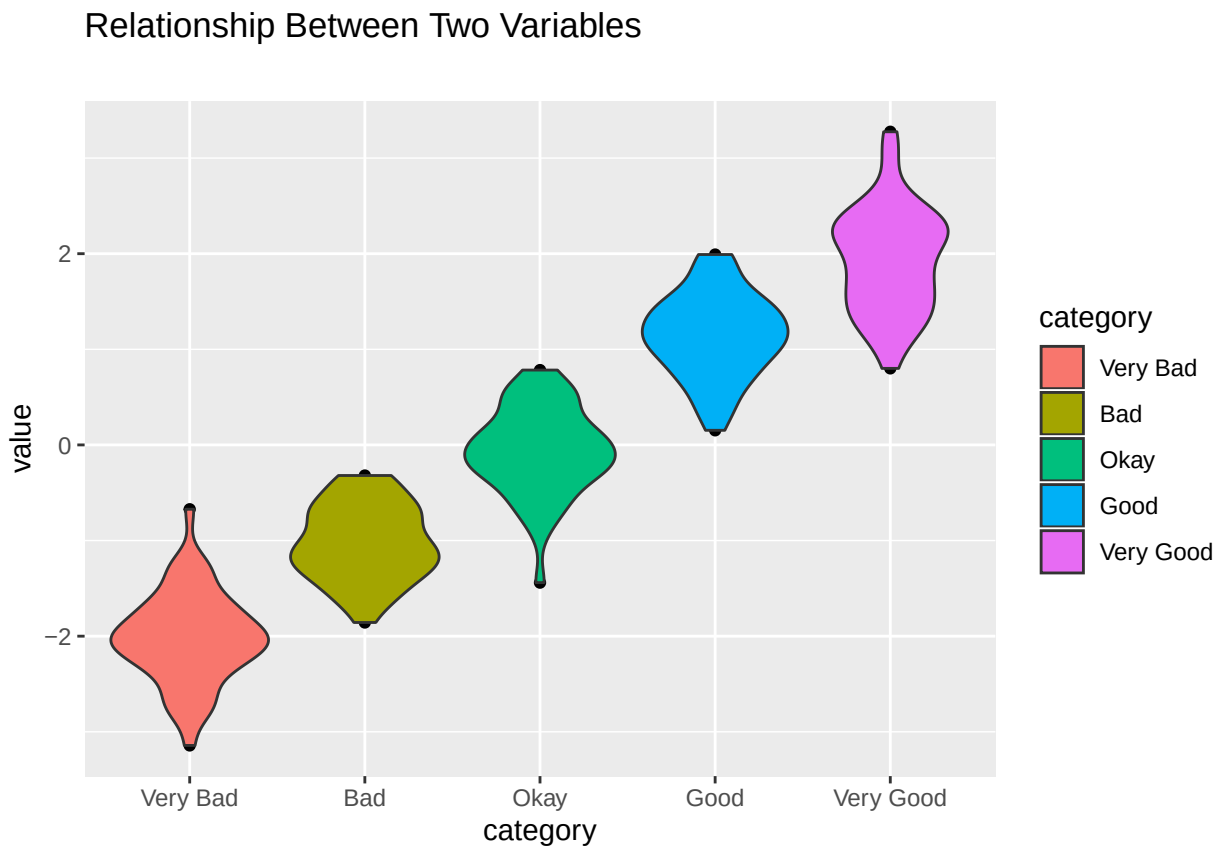
```
datset4 <- read_csv("data/datset4.csv")
```

```
## Parsed with column specification:
## cols(
##   category = col_character(),
##   value = col_double()
## )
```

```
dataset4 <- gather(datset4, category, value)
```

```r
dataset4 <-
  dataset4 %>%
  mutate(category = fct_relevel(category, "Very Bad", "Bad", "Okay", "Good", "Very Good"))

ggplot(dataset4, aes(x = category, y = value)) +
    geom_point() +
    stat_smooth(method = "lm",
        col = "#C42126",
        se = TRUE,
        size = 1) +
  labs(title= "Relationship Between Two Variables",
       subtitle="") +
   geom_violin(aes(fill=category))
```




Above is the relationship between the two variables in datset4. I used a boxplot at first but then decided to use geom violin to see the spread of the data. Pros of this can be seeing exactly how the data in each category can very, we dont see that very well with the box plots. A con of this is the reader can easily be confused by the shapes. Id they are unware of why they are shaped like this they could be put off by it. It was fun to make this graphic and playing around with the shapes and colors.